Testing for univariate Gaussian mixture in practice

Didier Chauveau, Bernard Garel, Sabine Mercier

Testing for univariate Gaussian mixture in practice*

Didier Chauveau†, Bernard Garel‡, Sabine Mercier§

November 30, 2017

We consider univariate Gaussian mixtures theory and applications, and particularly the problem of testing the null hypothesis of homogeneity (one component) against two components. Several approaches have been proposed in the literature during the last decades. We focus on two different techniques, one based on the Likelihood-Ratio Test, and another one based on estimation of the parameters of the mixture grounded on some specific adaptation of the well-known EM algorithm often called the EM-test. We aim to provide useful comparisons between different techniques, together with guidelines for practitioners in order to enable them to use theoretical advances for analysing actual data of realistic sample sizes. We finally illustrate these methods in an application to real data corresponding to the number of days between two events concerning ovarian response and lambing for ewes.

AMS 1991 subject classification: Primary 62F03; 62E20. Secondary 62F05.

Key words and phrases: Mixture models; likelihood ratio test; EM tests; Gaussian process.

1 Introduction

The aim of always producing the better model for our data can partly explain the present craze for probability distributions which can be written as a mixture. However there exist other important reasons which justify the increasing use of these models. Mixture models are able to help in many circumstances. Indeed, whenever a population is constituted of K homogeneous sub populations, a K-component mixture can be proposed as an attractive model for this population. Recently published books are entirely devoted to mixture of distributions. In particular, the books by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995), McLachlan and Krishnan (1997), McLachlan and Peel (2000), Böhning (2000), Frühwirth-Schnatter (2006), Schlattmann (2009) contributed to an already rich bibliography and became international references. We also have to stress that from the theoretical point of view, analysis of mixture involves many mathematical topics such as stochastic processes, asymptotic

*This research has been partially supported by the *Projet de Recherche d'Intérêt Régional* DURAREP2, reference 2011-00064290.

†Université d'Orléans, CNRS, MAPMO, UMR 7349, BP 6759, 45067 Orléans cedex 2, France. didier.chauveau@univ-orleans.fr

‡Université fédérale de Toulouse-Midi Pyrénées, ENSEEIHT, 2 rue Camichel, 31071 Toulouse Cédex 7, France; garel@math.univ-toulouse.fr

§Dpt Maths-Info, UFR Sciences Espace Société, Université Jean Jaurès, 5 allées A. Machado, 31058 Toulouse Cedex 9; mercier@univ-tlse2.fr

distribution, estimation and maximization for non regular models, Bayesian analysis and so on. Among the many problems raised by mixture we find the non-identifiability of parameters, the degeneracy of the Fisher information matrix around particular points or the non-differentiability with respect to another parametrization. Then it is not surprising that we obtain non-standard asymptotic distributions for testing problems.

Beginnings of mixture models go back to the 19th century mainly with the contributions by A. Quetelet, L.A. Bertillon, S. Newcomb and K. Pearson. Behind the writings of Quetelet (1846) we find the idea that a normal distribution can be generated by a great number of other normal distributions. Analysing the heights of 9002 conscripts, Bertillon (1874) and Bertillon (1876) noted that the graphical representation of these heights gave two modes which constituted a surprise. Then he claimed that this phenomenon was due to the presence of two distinct ethnic groups. The Figure he presented seems to be the first graphical representation leading to the assumption of a normal mixture. Newcomb (1882) and Newcomb (1886) addressed the problem of outliers in astronomical data. He observed that the tails of the distribution were fatter than the normal ones. He explained this non-normality by the combination of data with different scales and so, invented the contaminated normal distribution. Pearson (1894) analysed data that the zoologist Walter F. Weldon submitted to him, in particular crab forehead sizes. In his 1894 paper he graphically showed the evidence of a mixture. For him, to adjust a mixture of normal is equivalent to carve a skewed curve into two normal distributions. A way of doing so is to estimate five parameters from the five first moments. This is feasible because seeing that a mixture is a convex combination of densities, its moments are convex combination of the moments of these densities. Then Pearson found its famous ninth degree equation, a negative root of which is necessary to solve the problem. Pearson's contribution is generally thought of as the starting point of the analysis of mixtures. Then, during quite a while, the research on mixture models concentrated around improvements of this method of moments. Charlier (1906), Charlier and Wicksell (1924), Cohen (1967) and Holgersson & Jorner streamlined Pearson's claculations.

When the two variances are assumed equal, the problem consists in finding the negative root of a cubic. This problem was addressed by Pearson (1894), Charlier and Wicksell (1924), Rao (1948) and Cohen (1967). When the two means are assumed equal, the problem is still simpler and Gottschalk (1948), Cohen (1967), Gridgeman (1970) brought their contribution to it.

These contributions are related to the problem of estimation in mixture model. Another very important issue is the determination of the number of components of the mixture. Graphical procedures have been developed. Simple examination of the histogram can bring some information. A more elaborated method has been proposed by Bhattacharya (1967). The method starts from two statements. First the logarithm of a normal density is a concave quadratic in the variable, so that its derivative is linear with negative slope. Then, when there is a lot of data and the grouping imposed by the histogram is quite fine, the histogram heights are proportional to the density. Thus, a plot of first differences of the logarithms of the histogram frequencies should display a sequence of negatively sloped linear plots, one corresponding to each components. Roeder (1994) elaborated another type of graphical technic.

This concern has also been treated as a testing problem: a set of tests of k components against $k+1$ components, for instance using the likelihood ratio. After Wilks (1938), Chernoff (1954) gave the asymptotic distribution of the likelihood ratio test in the case of a regular model. Above, we gave a few reasons why mixtures do not belong to regular models. A few researchers began to privilege likelihood ratio for determining the number of components of a mixture and particularly the problem of testing H_0 : homogeneity ($k = 0$), against a two-component mixture.

Then conjectures and simulation results about the distribution of this likelihood ratio have been published.

Wolfe (1971) suggested that $(n-1-m)\lambda_n/n$ is approximately distributed as a χ^2_{2m}, where n is the sample size, λ_n is the usual likelihood ratio test statistic (LRTS) and m is the number of parameters which are different for the two components of the mixture. This gives a χ^2_2 distribution for a univariate normal mixture with equal variance and a χ^2_4 distribution for a normal mixture with different means and different variances.

In the case of univariate normal mixture with unknown but common variance, see Model 9 (M9) below, McLachlan (1987) and Thode et al. (1988) suggested that for $n \leq 1000$ the distribution of λ_n is close to a χ^2_2, the latter having a less heavy tail. In the case of a normal mixture with different means and different variances, see Model 8 (M8) below, McLachlan (1987) found that a χ^2_6 fits well for a sample size $n = 100$.

Hall and Stewart (1985) suggested using the restriction

$$\min_{1\leq i,j\leq 2}(\sigma_i/\sigma_j) \geq c \geq 0$$

and Feng and McCulloch (1996) used $\min(\sigma_1^2, \sigma_2^2) \geq c' \geq 0$. For $n = 100$, they found a distribution between a χ^2_4 and a χ^2_5 when $c' = 10^{-6}$ and between a χ^2_5 and a χ^2_6 when $c' = 10^{-10}$.

The preceding results, which are given without other restrictions on the parameters, rely on Monte Carlo simulations and concern essentially finite sample distributions. Indeed without assuming a bounded parameter the distribution tends to infinity in probability when n goes to infinity.

If we now consider asymptotic results, a first stage has been undertaken by Redner (1981). He proved that if W denotes a fixed neighbourhood of the set Γ corresponding to H_0 in the global parameter space, associated to the global model (1), then the probability that the maximum likelihood estimator (MLE) is found in W tends to one when n goes to infinity. Redner calls it convergence of the MLE in the topology of the quotient space obtained by collapsing Γ into a single point; see Ghosh and Sen (1985).

The first correct expression of the asymptotic distribution was given by Ghosh and Sen (1985) for a general mixture model with two components:

$$f(x; \pi, \mu_1, \mu_2) = (1-\pi)g(x, \mu_1) + \pi g(x, \mu_2), \tag{1}$$

where $\pi \in [0, 1]$ is the component weight and μ_1 (*resp.* μ_2) is the parameter of the first (*resp.* the second) component; π, μ_1 and μ_2 are unknown. The component density g in this model is general, assuming some regularity. First they need the assumption that the mixture parameters μ_1 and μ_2 belong to a bounded interval. Indeed, Hartigan (1985) proved that a statistic close to the LRTS for testing homogeneity against a Gaussian mixture of the means converges towards infinity in probability when n tends to infinity if the range of the unknown mean is unbounded. Bickel and Chernoff (1993) revisited this problem and showed that if the parameter set is unbounded, Hartigan's statistic approaches infinity with order $\log\log n$. Note that the equivalence between Hartigan's statistic and the LRTS, when H_0 is reduced to a single scalar parameter, has been proved not before quite recently by Liu and Shao (2003).

Ghosh and Sen also imposed a separation condition on the parameters of the mixture mainly in order to restore identifiability and to get an answer. They have assumed that $|\mu_2 - \mu_1| \geq c_0 > 0$. Therefore, under this constraint, H_0 is described by $\pi = 0$ or $\pi = 1$. Removing this

separation condition presented a real challenge and many statisticians offered a solution, for instance, Dacunha-Castelle and Gassiat (1997), Lemdani and Pons (1999), Liu and Shao (2003), Garel (2005). Garel (2001), Chen and Chen (2001), Garel and Goussanou (2002), Liu and Shao (2004) addressed specific mixtures in the Gaussian case. The LR approach has also being studied in a recent preprint (Maciejowska, 2013): the author proposes new hypothesis to test the homogeneity against two-component mixture model which allow to avoid the problem of identifiability.

When the parameter upon which relies the mixture is multivariate, the asymptotic distribution of the likelihood ratio is related to a Gaussian random field and the computation of percentile points becomes tricky or impossible. For instance, for a Gaussian mixture with unknown means and variances, the asymptotic distribution is a bivariate random field and, even in this case, it is not possible to get exact percentiles. That is why other tests or methods have been proposed in order to assess the number of components.

For a two-component mixture model, Chen et al. (2001), Chen et al. (2004) modified the LRTS and derived its limiting distribution. They used a penalized likelihood, with a penalty depending on the mixture proportion π. Li et al. (2009) proposed then an EM-test for homogeneity, that Chen and Li (2009) mentioned in the case of two normal mixture models. Li and Chen (2010) extend that to higher order (number of components) models, in situations where the distribution parameter is scalar (such as for Poisson or exponential mixtures). They test precisely a m_0-component mixture under the null hypothesis, vs. a mixture of order $m > m_0$. They claim for the limiting distribution under the null a mixture of $\delta_0, \chi^2(1), \ldots, \chi^2(m_0)$. Chen and Li (2011) propose a refined method for computing a tuning parameter in the penalty used in previous papers on this EM-test approach. Chen et al. (2012) then proposed an EM-test for testing the null hypothesis of some arbitrary fixed order under a finite mixture model.

Since these tests are intended to provide answers to end users, and since they require some sort of heavy computations, the availability of public codes is important. Several versions of these EM-tests have been made publicly available in the recent MixtureInf package (Li et al., 2016) for the R statistical software (R Core Team, 2016). Two successive versions of this package have been proposed, a first version (1.0-1) in 2015 and the most recent update (version 1.1, March 2016). In this current version, the function `emtest.norm` is dedicated to the test of the order of a normal mixture model. The linked references are precisely Chen and Li (2009), that was limited to the homogeneous null model vs. a two-component mixture, and Chen et al. (2012) which generalize this EM-test approach to a mixture of arbitrary order under the null hypothesis.

In this paper, we investigate some of these methods from the practitioner point of view, i.e. practical applicability and power that can be expected. We focus on two different techniques, the LRT and the EM-test based on the EM algorithm. We aim to provide useful comparisons between these techniques in several (more or less general) Gaussian mixture models, and guidelines for practitioners in order to enable them to use these methods for analysing actual data of realistic sample sizes. We have also developed numerical procedures for the EM approach for some of the models that are not available in the MixtureInf package, such as Model 2 and 9 (M2 and M9) in Table 1.

Table 1 describes the models studied from the perspective of testing homogeneity vs. mixture in the literature, where the labels are borrowed from previous literature. The models we actually investigate in this paper are in boldface.

On the computational side, since there exists (up to our knowledge) no public codes for the LRT approach, we develop numerical procedures that will be publicly available in an upcoming

Table 1: Description of the models studied from the perspective of testing homogeneity vs. mixture in the literature (models investigated in this paper are in boldface).

Models	H_0	H_1
1- Contaminated Models		
(M1)	$g(\alpha_0)$ $\alpha \in A \subset R$	$(1-\pi)g(\alpha_0) + \pi g(\alpha)$
(M2)	$\mathcal{N}(0,1)$	$(1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(\mu,1)$ $\mu \in A \subset R$
(M3)	$\mathcal{N}(0,1)$	$(1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(0,\sigma^2)$ $\sigma^2 \in [a,A] \subset]0,2[$
(M4)	$\mathcal{N}(0,1)$	$(1-\pi_1-\pi_2)\mathcal{N}(0,1)+$ $\pi_1\mathcal{N}(\mu_1,1) + \pi_2\mathcal{N}(\mu_2,1)$
2- One population against two		
(M5)	$f(x,\alpha)$ $\alpha \in A \subset R$	$(1-\pi)f(x,\alpha_1) + \pi f(x,\alpha_2)$
(M6)	$\mathcal{N}(\mu,1)$ $\mu \in A \subset R$	$(1-\pi)\mathcal{N}(\mu_1,1) + \pi\mathcal{N}(\mu_2,1)$
(7)	$\mathcal{N}(0,\sigma^2)$	$(1-\pi)\mathcal{N}(0,\sigma_1^2) + \pi\mathcal{N}(0,\sigma_2^2)$ $\sigma_2^2 < 2\sigma_1^2$
(M8)	$\mathcal{N}(\mu,\sigma^2)$	$(1-\pi)\mathcal{N}(\mu_1,\sigma_1^2) + \pi\mathcal{N}(\mu_2,\sigma_2^2)$
3- Presence of a structural parameter		
(M9)	$\mathcal{N}(\mu,\sigma^2)$	$(1-\pi)\mathcal{N}(\mu_1,\sigma^2) + \pi\mathcal{N}(\mu_2,\sigma^2)$ σ^2 unknown
(M10)	$\mathcal{N}(\mu,\sigma^2)$	$(1-\pi)\mathcal{N}(\mu,\sigma_1^2) + \pi\mathcal{N}(\mu,\sigma_2^2)$ μ unknown

version of the mixtools package (Benaglia et al., 2009) for the R statistical software (R Core Team, 2016). We compare these LRT codes with some of the codes proposed in the MixtureInf package (Li et al., 2016) (see above).

The rest of the paper is organized as follows: Sections 2 to 5 are dedicated to analyses for models (M2), (M6), (M8) and (M9) respectively. Section 6 presents some applications based on actual data collected for a research project from the French National Institute for Agricultural Research (INRA)[1]: we study the so-called "ram effect" on data corresponding to number of days between two events concerning ovarian response and lambing for ewes of several kinds. For such data, a mixture is sometimes suspected for biological reasons, but with not great evidence coming from the empirical distribution. The Discussion section 7 summarizes the results and derive some practical suggestions for the end users.

[1] *Projet de Recherche d'Intérêt Régional* DURAREP2.

2 Model 2

This simplest model is the standard normal $\mathcal{N}(0,1)$ contaminated by a normal distribution shifted by a mean μ,

$$H_0 : \mathcal{N}(0,1) \quad \text{vs.} \quad H_1 : (1-\pi)\mathcal{N}(0,1) + \pi\mathcal{N}(\mu,1). \tag{2}$$

This model is obviously rather artificial from a practical point of view. We start our study with it since it is the first time, up to our knowledge, that the actual quantiles for finite (realistic) sample size n are compared to the asymptotic quantiles obtained by Garel (2001). The Likelihood Ratio Test statistic (LRT) proposed by Garel (2001), Theorem 2.1, is

$$\lambda_n = \sup_{\mu\in[-a,a]\setminus\{0\}} T_n^2(\mu)\, \mathbb{I}_{\{T_n(\mu)\geq 0\}} \tag{3}$$

where

$$T_n(\mu) = \frac{1}{\sqrt{n(e^{\mu^2}-1)}} \sum_{i=1}^{n} \left[e^{(X_i\mu - \mu^2/2)} - 1 \right], \quad \mu \neq 0,$$

and

$$\lim_{\mu\to 0} T_n^2(\mu) = n\bar{X}^2, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Our purpose in this section is to compute Monte-Carlo quantiles of this test statistic for realistic n (and up to "asymptotic" sample sizes), and evaluate the asymptotic behaviour w.r.t. the theoretical results. This allows us to evaluate the power of this LR-Test using these Monte-Carlo or asymptotic quantiles. Finally, we are comparing it with an EM-test we derived using the methodology in Chen and Li (2009). Note that this particular model has not been handled by these authors.

2.1 Model 2 Monte-Carlo simulation for quantiles

2.1.1 Quantiles for the LR Test

For computing the statistic, we have to define a suitable compact $[-a,a]$. Following Garel (2001), we first tried values $a \in \{1, 2.5, 5\}$. The test statistic λ_n is easy to compute, the supremum over $\mu \in [-a,a] \setminus \{0\}$ being obtained by discretizing the interval in $k = 100$ or $k = 200$ steps. Fig. 1 shows some typical behaviour of $\mu \mapsto T_n^2(\mu)\mathbb{I}_{\{T_n(\mu)\geq 0\}}$, for which we choose $a = 2.5$ that allows to see the global behaviour of the statistic. In particular the discontinuity in 0 where the statistic jumps to its opposite value is visible.

Fig. 2 shows the comparison between asymptotic, previously published quantiles, and Monte-Carlo quantiles computed from a large-scale experiment with $10,000$ replications, and several sample sizes from $n = 100$ up to $n = 100,000$. This experiment shows that the convergence is rather slow, but happened for $a = 1$, whereas the usage of the asymptotic quantile is questionable in cases where $a = 2.5$ or larger is used. This study hence suggests to use the Monte-Carlo quantiles in practice for any realistic (hence small) n. This very slow convergence is somehow in accordance with the rate in $\log(\log n)$ claimed by other authors as, eg, Bickel and Chernoff (1993). Note that using $a > 2.5$ is not realistic in practice since, for contamination mean μ so distant from 0, the mixture structure becomes visible just looking at an histogram of the data.

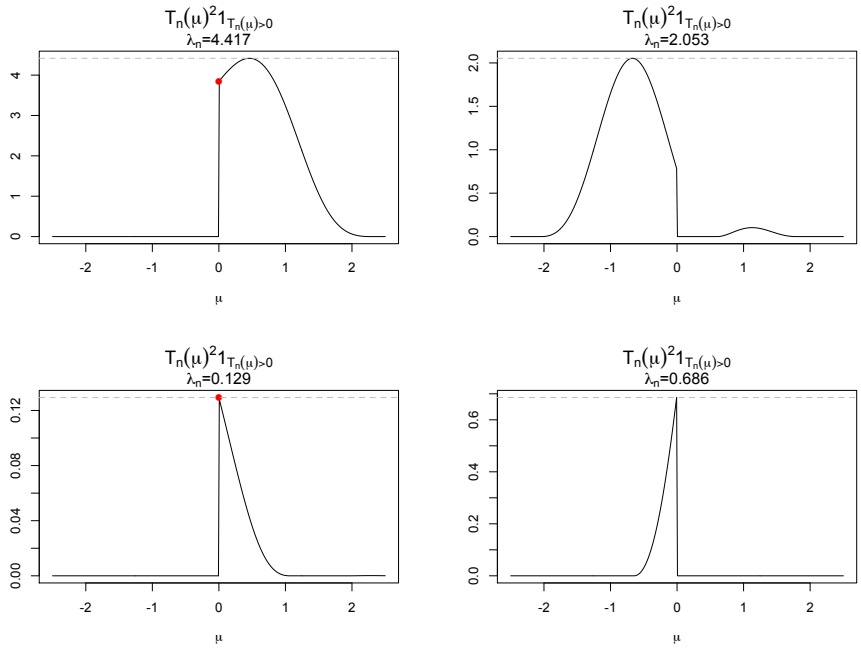

Figure 1: Some typical behaviours of $\mu \mapsto T_n^2(\mu)\mathbb{I}_{\{T_n(\mu)\geq 0\}}$ for $\mu \in [-2.5, +2.5]$ and simulated samples of size $n = 1000$ under H_0 for (M2). The red dot is the limiting behaviour $T_n^2(0)$, when $T_n(0) \geq 0$.

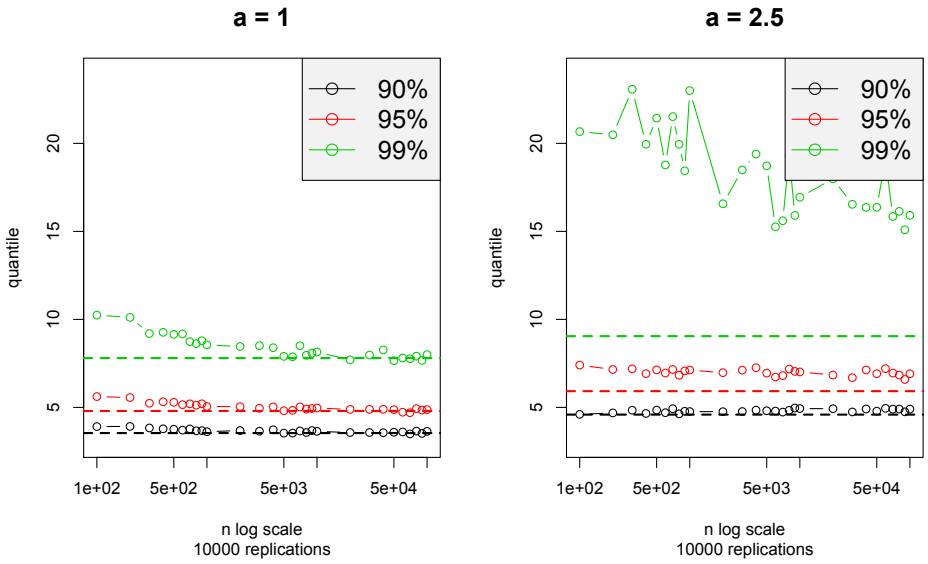

Figure 2: (M2): Empirical quantiles for λ_n based on 10,000 Monte-Carlo replications vs. Asymptotic theoretical values from Garel (2013) (dotted lines) for three levels (90%, 95% et 99%) and two compact sets with $a = 1$ (left) and $a = 2.5$ (right). The x axis is in log scale, for $100 \leq n \leq 100,000$.

2.1.2 Quantiles for the EM-test

We present here the quantiles calculated from Monte-Carlo simulation for the statistic $EM_n^{(k)}$ of the EM-test proposed in Chen and Li (2009). (M2) corresponds to Exemple 2 in Li et al. (2009) but it seems that they do not provide the asymptotic distribution. We did not find a definition and implementation of the EM-test for this model in the package MixtureInf (Li et al., 2016) proposed by these authors, so we defined our implementation. These quantiles are computed from experiments with 10,000 replications of size n. Each experiment have been computed three times to evaluate the accuracy of the number of replications.

We choose $K = 3$ for the number of iterations in the EM algorithm and (0.1, 0.3, 0.5) for initial values for π as proposed in, e.g., Chen and Li (2009). The maximization of the initial step is done using the R function `optimize()` for this simple case. We use the penalization proposed in Chen and Li (2009), $p(\pi) = \log(1 - |1 - \pi|)$. Our results are in Table 2.

Table 2: Mean (and standard deviations) of the quantiles of the statistic $EM_n^{(K)}$ for different probabilities and different values of n.

n/α	90%	95%	99%
100	2.81 ($5.97\ 10^{-2}$)	3.97 ($1.77\ 10^{-2}$)	6.6 ($8.78\ 10^{-2}$)
200	2.75 ($5.8\ 10^{-2}$)	3.94 ($8.71\ 10^{-2}$)	6.89 ($22.8\ 10^{-2}$)
500	2.75 ($3.47\ 10^{-2}$)	3.93 ($3.98\ 10^{-2}$)	6.75 ($7.22\ 10^{-2}$)
1000	2.68 ($0.47\ 10^{-2}$)	3.83 ($1.90\ 10^{-2}$)	6.48 ($17.6\ 10^{-2}$)
5000	2.73 ($4.63\ 10^{-2}$)	3.90 ($10\ 10^{-2}$)	6.56 ($17.6\ 10^{-2}$)
10^4	2.71 ($0.54\ 10^{-2}$)	3.87 ($2.30\ 10^{-2}$)	6.65 ($10.3\ 10^{-2}$)
$\chi^2(1)$	2.71	3.84	6.63

From our Monte-Carlo experiment, a $\chi^2(1)$ limit distribution for the EM-test statistic $EM_n^{(K)}$ for (M2) seems valid, even though the convergence appears to be very slow.

2.2 Model 2 power evaluation

We have simulated under H_1 the mixture:

$$X \sim (1 - \pi)\mathcal{N}(0, 1) + \pi\mathcal{N}(\mu, 1), \tag{4}$$

with parameters $\mu = 0.7$ and $\pi = 0.3$, already used in (Garel, 2001). These settings result in a severely overlapping mixture, as illustrated in Fig. 3. Our motivation for choosing such a non-obvious mixture model is based on the idea that, if a simple histogram of the data already reveals a multi-modal distribution, then the test itself is not needed. Fig. 4 shows that the estimated power of the LRT for model (4) is very good, even for small sample sizes, considering the difficulty of this severely overlapping mixture. However, one should be aware that this good behaviour is also a consequence of the simplicity of the model, with very few unknown parameters and a completely known distribution under H_0. Our implementation of the EM-test for M2 shows a comparable power.

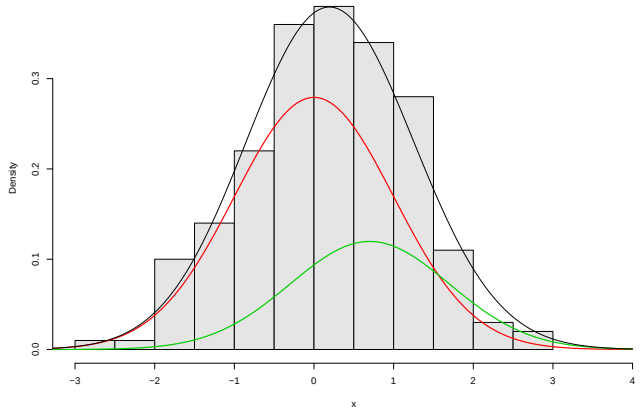

Figure 3: Typical empirical distribution of a $n = 200$ sample from the mixture model (4), with the true distributions for the two components (red, green) and the mixture (black).

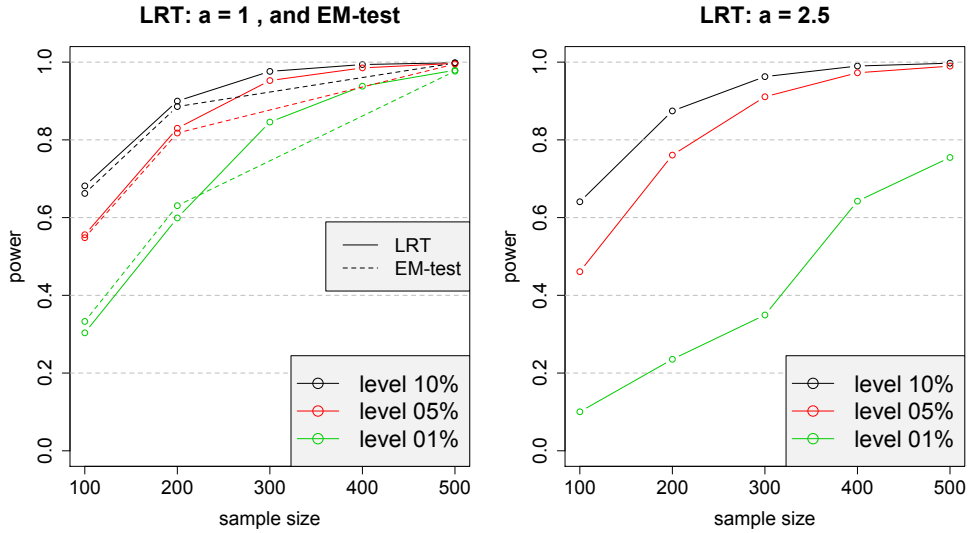

Figure 4: Power estimates for M2. LRT: using the quantiles obtained by Monte-Carlo as in Fig.2, for $a = 1$ (left) and $a = 2.5$ (right), $10,000$ replications. EM-test (left): for $n = 100, 200$ and 500 replications.

3 Model 6

This model corresponds to a location mixture with same and known variance, set to 1 without loss of generality.

$$(M6) \quad H_0 : \mathcal{N}(\mu_0, 1) \quad \text{vs.} \quad H_1 : (1-\pi)\mathcal{N}(\mu_1, 1) + \pi\mathcal{N}(\mu_2, 1).$$

A LRT statistic has been proposed for this model in Garel (2001). Its formulation is in its principle similar to λ_n for (M2), Equation (3), and we refer to Garel (2001) for its detailed

definition, that also includes a definition of $T_n(\mu)$. The behavior of this statistic has been studied only asymptotically ($n = \infty$). The drawback of this approach is that the definition of the statistic includes the knowledge of the true value of μ_0 under H_0, that makes sense asymptotically but causes problem in practice for actual data, where the mean have to be estimated from other "reference" data corresponding to the null hypothesis H_0, prior to use the test. Figure 5 shows some typical behaviour of $\mu \mapsto T_n^2(\mu)\mathbb{I}_{\{T_n(\mu)\geq 0\}}$ for $a = 2.5$ and $\mu_0 = 0$. As for (M2), we have compared the asymptotic quantiles of the LRT statistic with actual quantiles obtained by a Monte-Carlo experiment. In the asymptotic framework (μ_0 known), Fig. 6 shows the slow convergence of the actual quantiles toward the asymptotic ones.

An EM-test for (M6) is provided in Li et al. (2009), since it is a special case of their general result for a scalar parameter θ, and a general density f satisfying regularity assumptions. Their theorem 2 says that the limiting distribution (in n) of the test statistic $EM_n^{(K)}$ for any K is

$$EM_n^{(K)} \to 0.5\,\delta_0 + 0.5\,\chi^2(1).$$

This EM-test has been implemented in the package MixtureInf (Li et al., 2016), in the two versions we tested. Note that the code was provided by the function `emtest.norm` in their first version, and has been renamed `emtest.norm0` in the last available version we tried (2016). We have checked numerically that the statistic actually converges to the claimed distribution for realistic sample sizes. However, we noticed that the discrete part $0.5\,\delta_0$ comes in their code from the fact that any negative value of the statistic are simply replaced by 0, which is a procedure not in accordance with the theory, and for which we have no explanations. All the experiments showed here have been done with the last package version 1.0-1.

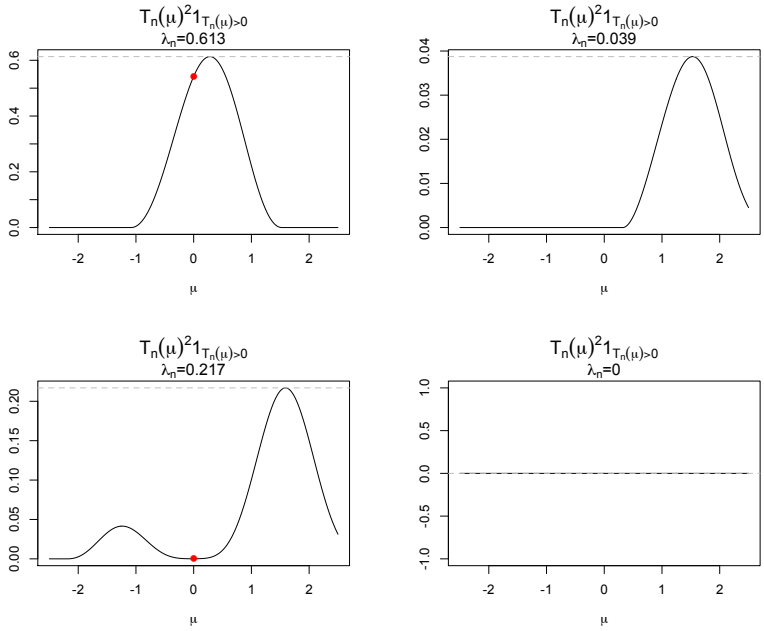

Figure 5: Some typical behaviour of $\mu \mapsto T_n^2(\mu)\mathbb{I}_{\{T_n(\mu)\geq 0\}}$ for simulated samples of size $n = 1000$ under $H_0 : \mathcal{N}(0, 1)$ for (M6). The red dot is the limiting behaviour $T_n^2(0)$, when $T_n(0) \geq 0$. The statistic can be null, when $T_n(\mu) < 0$ over all the compact set, as illustrated in the bottom right panel.

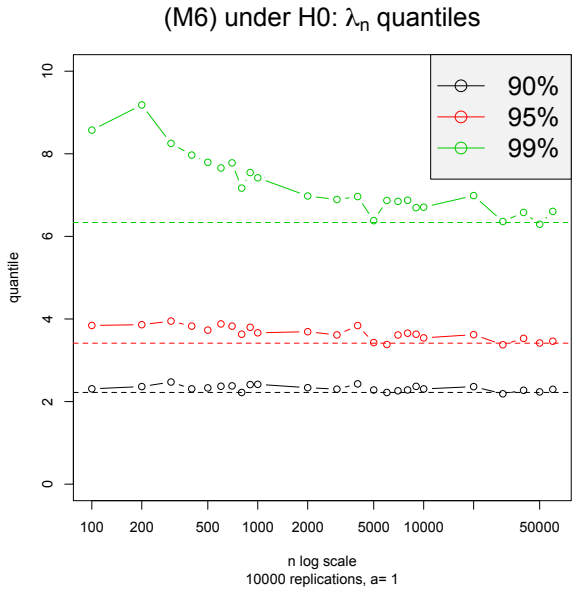

Figure 6: (M6) Monte-Carlo quantiles of the LRT statistic under H_0 based on 10,000 Monte-Carlo replications, assuming μ_0 known, compared with asymptotic values from Garel (2001), with supremum over $[0, a]$ and $a = 1$. The n axis is in log scale.

3.1 Model 6 power evaluation

As mentioned above, we have used the function `emtest.norm0()` from the last available update of the R package MixtureInf that implements the EM test for (M6) in our Monte-Carlo simulation for power evaluation, with the same model under H_1 as for M2, equation (4), and same sample sizes and test levels. We have also compared these results with a similar experiment using the LRT approach and test statistic from Garel (2001), but remembering that this statistic for (M6) requires the knowledge of the true value of μ under H_0. The results in Fig. 7, left, shows that the power here is weaker than the LRT in (M6) (Fig. 7, right) and the LRT in (M2) studied previously. However, remember that the comparison with the power of the LRT test in (M6) is questionable, since for the EM test the model have to estimate μ under the null hypothesis, whether it is provided in the LRT asymptotic form. Our advise is then that the LRT is a better option if the mean under homogeneity (H_0) is known from previous experiments, expert prior information, or as a standard from some area of expertise. Note that the comparison with the LRT in (M2) is also not meaningful since (M2) is an even simpler model with all parameters known under H_0. Note that the different number of replications between the LRT and the EM-test is only motivated by the computing time required for the latter.

4 Model 8

This model is actually the general one, in particular more general than (M9) that will be considered later, since all the parameters are unknown and unconstrained under both H_0 and H_1.

$$H_0 : X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{vs.} \quad H_1 : X \sim \lambda\mathcal{N}(\mu_1, \sigma_1^2) + (1-\lambda)\mathcal{N}(\mu_2, \sigma_2^2),$$

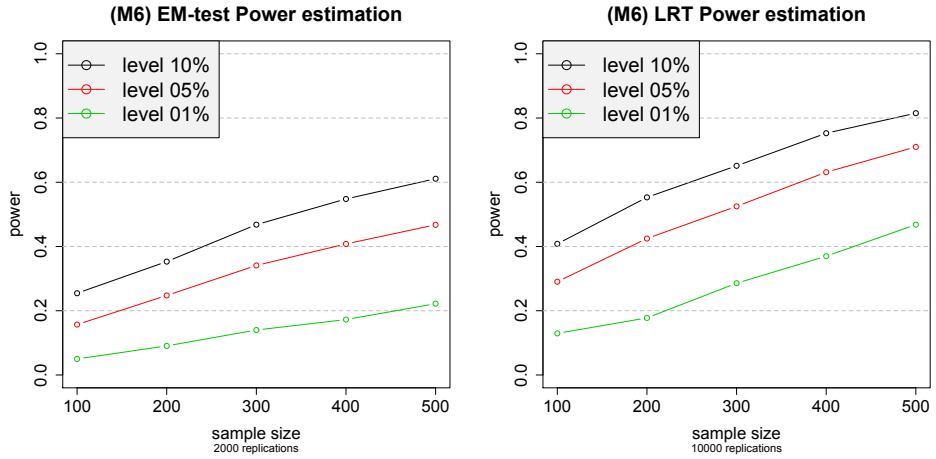

Figure 7: Monte-Carlo estimation of the power for (M6), parameters as in equation (4). *Left:* EM-test from Li et al. (2009), using 2000 replications; *Right:* LRT approach as in Garel (2001) using quantile obtained by Monte-Carlo, but using the asymptotic definition of the statistic, i.e. μ_0 known; 10,000 replications.

i.e. the test for H_0 : "Gaussian distribution" vs. H_1 : "Two-component Gaussian mixture". For this model, the LRT-based strategy only proposed a test statistic as a conjecture (Garel, 2001). Chen and Li (2009) Section 3 handles this model with the EM-test, and claim that the limiting distribution is $\chi^2(2)$. The code for the EM-test in (M8) is provided by the function `emtest.norm` in the last available update of the R package MixtureInf (Li et al., 2016). This code also handles these authors recent extensions to higher order tests, namely a mixture with m_0 components for the null hypothesis versus a mixture with $m > m_0$ components, see Chen et al. (2012).

We have experimented this EM-test strategy under H_0 first. The announced limiting distribution of the M_n statistic as a $\chi^2(2)$ is partially verified in practice, in the sense that, under $H_0 : \mathcal{N}(0, 1)$, we always observed a small percentage (between 3% and 5%) of negative values. If we remove these negative values, then the empirical distribution is reasonably fitted by a $\chi^2(2)$.

We have also evaluated the power of the EM-test for (M8) using a Monte-Carlo experiment as before, for the same mixture model as in (M2) and (M6), i.e. with parameters as in equation (4). The results are rather disappointing, in comparison with, e.g., (M6), even if the statistical problem is obviously more difficult here. Figure 8, left, shows the empirical power using the same settings as before. It is clear that our model with means 0 and 0.7 is too difficult for this test to capture the (severely overlapping) mixture structure. Actually, the EM-test power here is approximately equal to its level.

We have thus conducted a second experiment, where the power is estimated when the mean of the second component increases, i.e. the mixture becomes more and more easy to detect, for the same range of sample sizes. Results are displayed in Figure 8 (right).

Note that we ran the same experiment initially with the previous version (1.0) of the MixtureInf package, which returns results similar to Fig. 8 (Left) for our model under H_1, and slightly worse than Fig. 8 (Right) when estimating the power as a function of the increasing μ_2. For instance in the case where $n = 100$, and for the most obvious mixture with $\mu_2 = 3$, the power of the previous version was about 50% instead of the 80% showed here, for a level 1%.

Our advice to users for this general model is that a good power can only be obtained for

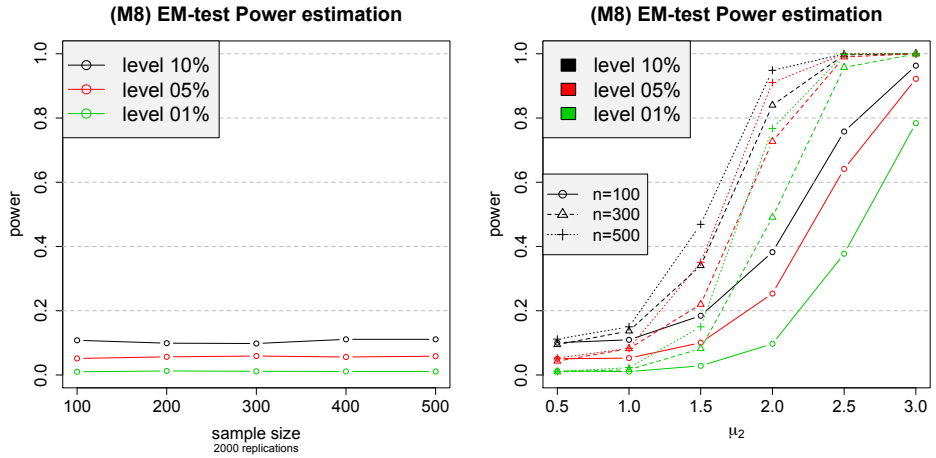

Figure 8: Monte-Carlo estimation of the power for (M8) based on 2000 replications: *Left*, parameters as in equation (4), using the EM test from Chen Li & Marriot (2009) and Chen et al. (2012) referred to in the MixtureInf package. *Right*: Power as a function of the mean μ_2 of the second component.

"separated enough" models. Fig. 9 illustrates for instance the type of true and empirical distributions one can expect, to achieve a power $\geq 90\%$ with a sample of size $n = 300$. It shows in particular that in this case the empirical distribution can be at least severely skewed, and often even bimodal, so that H_0 is not reasonable: the decision criterion provided by the test seems thus limited in practice.

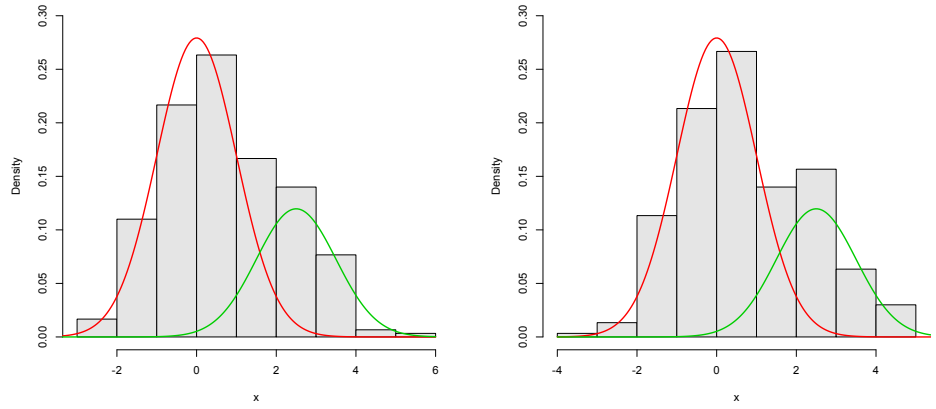

Figure 9: Mixture (M8) under H_1 for $\mu_2 = 2.5$, and two examples for samples of size $n = 300$.

In view of our results for the power of the EM-test in the case of a homogeneous null model, and since this model has been extended in Chen et al. (2012) and in the MixtureInf package (Li et al., 2016), to higher order models i.e. for $H_0 : m_0$-component mixture vs. $H_1 : m > m_0$-component mixture, we tried to re-run some of the experiments provided in Chen et al. (2012), Section 4 (simulation study). We reproduced the experiments given by these authors in their Table 4, bottom panel, corresponding to $H_0 : m_0 = 2$ vs. four instances of $H_1 : m = 4$, and their Table 6, bottom panel, corresponding to $H_0 : m_0 = 3$ vs. four instances of $H_1 : m = 5$. Results

are summarized in our Table 3. We estimate the powers based on 5000 replications (instead of 1000 in Chen et al. (2012)), to achieve more precise estimates. The results are sometimes surprising in comparison to the original estimates, without a clear explanation.

Table 3: Powers of EM test at the 5% level for two models (M8) and four alternatives for each, compared with results from Chen et al. (2012) showed in (·) for the EM-test with 3 iterations; estimates based on 5000 replications.

	H_1	$n = 200$	$n = 400$
$m0 = 2$ vs. $m = 4$ (Table 4)	1	16.2 (20.0)	43.9 (44.1)
	2	33.2 (33.5)	67.6 (70.2)
	3	99.1 (40.5)	100 (60.2)
	4	100 (100)	100 (100)
$m0 = 3$ vs. $m = 5$ (Table 6)	1	13.0 (10.4)	25.9 (28.4)
	2	41.8 (40.7)	79.7 (84.6)
	3	42.7 (44.8)	78.6 (83)
	4	80.1 (82.3)	99.2 (99.5)

5 Model 9

This model is as before a mixture on the mean, variance unknown but common (sometimes called the structural parameter). It is thus a generalization of (M6), but less general than M8.

$$H_0 : \mathcal{N}(\mu_0, \sigma^2) \quad \text{vs.} \quad H_1 : (1-\pi)\mathcal{N}(\mu_1, \sigma^2) + \pi\mathcal{N}(\mu_2, \sigma^2). \tag{5}$$

A LRT statistic is available for this model in Garel (2001) section 3.3, but it is again asymptotic in the sense that the statistic includes the knowledge of both the mean μ_0 under H_0 (as for M6) and of the structural parameter σ^2. We thus do not investigate its behaviour in practice here.

This model is also studied in Chen and Li (2009), Section 2, where the EM-test limiting distribution is given by their Theorem 2: for any fixed number of iterations K of the EM algorithm using a penalized log-likelihood,

$$\mathbb{P}(EM_n^{(K)} \leq x) \to F(x - \Delta)[0.5 + 0.5F(x)] \quad \text{as } n \to \infty,$$

where $F(\cdot)$ is the cdf of the $\chi^2(1)$ distribution, and Δ is a negative but fixed constant that depends only on the penalty $p(\pi)$ and the π_j's chosen in the initialization procedure (Chen and Li, 2009). This distribution has its support in (Δ, ∞) and

$$\mathbb{P}(EM_n^{(K)} \leq 0) \to 0.5F(-\Delta) \quad \text{for } \Delta < 0.$$

In particular with the settings from Chen and Li (2009),

$$p(\pi) = \log(1 - |1 - 2\pi|), \quad \Delta = 2 \max_{\pi \in \{0.1, \ldots, 0.4\}} (p(\pi) - p(0.5)) \approx -0.446,$$

so that it gives $\mathbb{P}(EM_n^{(K)} \leq 0) \approx 0.248$.

We did not find any implementation of the EM-test in this case in the MixtureInf package, even though the algorithm and the associated penalty functions (for $p(\pi)$ and $p(\sigma)$) are fully described in Chen and Li (2009) section 2. Hence we develop our implementation of this EM-test to compare it with (M8). We were unable to clearly validate the asymptotic distribution under H_0, essentially because we observed a higher estimated value for $\mathbb{P}(EM_n^{(K)} \leq 0)$, even for large samples up to $n = 2000$. Our estimated type I error (Table 4) also did not exactly match the simulations provided by Chen and Li (2009) Table 1. We finally also apply the EM-test for (M9) with the same settings already used for (M8), as in Fig.8 (right). Results are in Fig.10 and show a slightly better power, which is expected since the model is simpler, but again good power is associated to somehow "easy to detect" mixtures.

Table 4: Estimated type I error for sample size $n = 200$, $K = 3$ and $\pi \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$; 20,000 replications.

Level	10%	5%	1%
Estimates	7.81	4.16	0.99

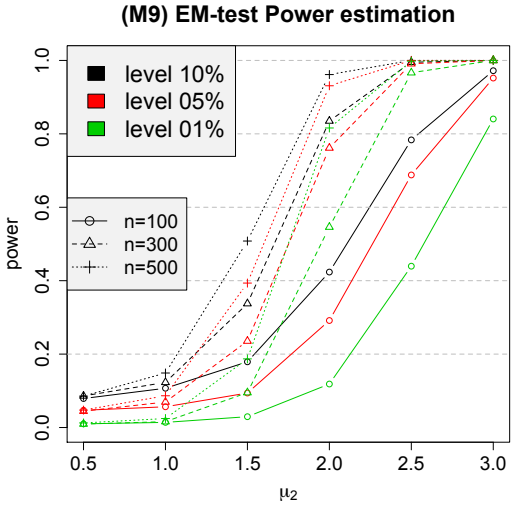

Figure 10: Monte-Carlo estimation of the power of the EM-test for (M9) based on our implementation and 2000 replications, as a function of the mean μ_2 of the second component.

A note about CPU time and computing efficiency: our code of the EM-test for (M9) uses C calls for computing the E-step, borrowed from the mixtools package (Benaglia et al., 2009). We noticed that for the Monte-Carlo simulations used in power estimations showed in Figures 8 (right) and 10, our code is approximately 30 times faster than the EM-test for (M8) provided by the MixtureInf package.

6 An application to real data

To illustrate the application of some of the previous models, quantiles and power to actual data of moderate sample sizes, we have applied it to data collected for a research project from the

French National Institute for Agricultural Research (INRA)[2]. Briefly, these quantitative data correspond to the number of days between two events concerning ovarian response and lambing for ewes of several kinds (races), coming from actual farms or experimental situations, and years. The purpose is the study of the so-called "ram effect" (or male effect). A mixture is sometimes suspected for the distributions of these datasets, and the empirical distributions are not always giving evidence of it. Hence the test for homogeneity comes as a natural technique to asses mixture or homogeneous population. The conclusions of the tests have important consequences for the biological application.

We focus here on a dataset of $n = 660$ observations for a selected race (*Romane*), location (*Sapinière*) and year (2009). Fig. 11 shows the empirical distribution of the data, which does not look very well bell-shaped, but is not obviously bimodal in the sense expected from the biological context (a mode around 150 days and another mode about 6 days later). We have first added to this plot a single normal fit (i.e., assuming homogeneity), and a two component Gaussian mixture fit. All the mixture model fits in the figures were done using the normalmixEM() function of the Benaglia et al. (2009) mixtools package for the R statistical software R Core Team (2016). The initialization of the EM algorithm is data-driven there, based on an initial k-means clustering of the data.

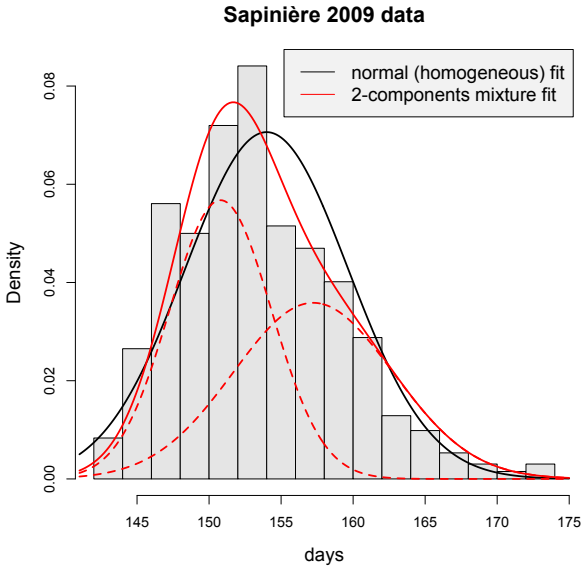

Figure 11: Homogeneous and 2-components mixture fits for the data *Romane Sapinière 2009*.

(M6) can reasonably be applied to these data, assuming equal variance among the components, and hence standardizing the data to bring it back to variance 1 as in (M6) set-up (both for LRT and EM-test approaches). The EM-test returns a p-value of 1.3% signifying rejection of homogeneity. The LRT approach implies, as detailed in Section 3, the knowledge of the mean μ_0 under H_0. The current practice is to estimate this mean from the data, even though this is not theoretically valid since the plug-in of this estimate adds some noise to the statistic's asymptotic distribution. As explained in Section 3 and illustrated in Fig. 6, it is preferable to use the Monte-Carlo quantiles instead of the less conservative asymptotic quantiles. Doing this here returns the results in Table 5.

[2] *Projet de Recherche d'Intérêt Régional* DURAREP2.

Table 5: Results for the LRT and EM-test using (M6), Monte-Carlo quantiles and plug-in estimate of μ_0 computed on the standardized data.

Method	90%	95%	99%	L_n	p-value
LRT, $a = 1$	2.38	3.83	7.78	10.82	< 0.01
LRT, $a = 2.5$	2.86	4.66	11.63	14.71	< 0.01
EM-test					0.013

The LRT concludes rejection (i.e. existence of a mixture) even at level 1%, thus in accordance with the EM-test under the same (M6) set-up. A study of the detailed plot of $\mu \mapsto T_n^2(\mu)\mathbb{I}_{\{T_n(\mu)\geq 0\}}$ as in Fig. 5 for these data shows a behaviour similar to Fig. 5 top-right panel, so that increasing a does not change the result. The exact p-value, clearly smaller than 1%, has not been computed here, but this could be done using the Monte-Carlo experiment presented in Section 3.

(M8) can also be used straightforwardly in this case, but only (among the various approaches we have detailed) using the EM-test approach. Note that applying the EM-test using emtest.norm2() on this dataset with the pre-2016 version of the MixtureInf package return a negative value for the statistic, and consequently a p-value of 1, not satisfactory in practice in view of the empirical distribution. The updated version (2016) returns a p-value of 5.10^{-9} indicating clear rejection, in accordance with the results based on (M6) given in Table 5.

7 Discussion

7.1 A brief summary of our results and observations

M2 Results on the LRT statistic distribution is available (Garel, 2001). The comparison between the non asymptotic quantiles vs. the asymptotic ones shows the good power even for a severely overlapping and non obvious mixture. To our knowledge, there is no EM-test available in the literature. Our derivation of an EM-test theory and implementation indicates a limiting distribution under the null hypothesis

$$EM_n^{(K)} \to \chi^2(1),$$

and power similar to the results from the LRT test. Note that the convergence to the $\chi^2(1)$ distribution seems rather very slow.

M6 The LRT approach for (M6) gives an asymptotic results on the statistic distribution, but the expression of the test statistic includes the value of the true μ_0 under H_0. Quantiles and power are evaluated in our work, in this set-up. For the EM-test approach, Li et al. (2009) Theorem 2, propose the following limit distribution

$$EM_n^{(K)} \to 0.5\,\delta_0 + 0.5\,\chi^2(1).$$

Note that a code is available in the MixtureInf package. We however noticed that values corresponding to the discrete part δ_0 are forced in the code, where negative values for $EM_n^{(K)}$ are

checked and simply replaces by 0. This is also the case for the last 2016 MixtureInf package update. The limit distribution under H_0 is recovered from our Monte-Carlo investigation, except the convergence to the weight 0.5 even for large n. We observe that the power in this model is weaker than the LRT. Our advise is then that the LRT is a better option if the mean under homogeneity (H_0) is known from previous experiments, expert prior information, or as a standard from some area of expertise.

M8 In that model, a conjecture is proposed for the statistic distribution but it has not been investigated. In Chen and Li (2009) Section 3, a limiting distribution is given for the EM-test statistic

$$EM_n^{(K)} \to \chi^2(2).$$

Code for this case is available in the MixtureInf package. We tested it and observed negative values for $EM_n^{(K)}$ in the previous package version (1.0-1), that gave obviously wrong results (p-value of 1 i.e. no rejection even for obvious mixtures like, e.g., the Old Faithful geyser waiting time data). The last MixtureInf package version (1.1 updated in March 2016) successfully corrects that problem. Following our investigations, the message we deliver to potential users is that the power can be rather weak for datasets from non-obvious mixture, i.e. non obviously bimodal (or multi-modal) empirical distributions. (M8) clearly deserves more study on actual datasets.

M9 Results on the LRT are available but as for M6, only asymptotically in the sense that the computation of the test statistic requires the knowledge of both the mean μ_0 under H_0 and the structural parameter σ^2. It is thus not investigated here. The EM-test has been studied in Chen and Li (2009) section 2 and their Theorem 2 gives an asymptotic behaviour of the cdf of the statistic test $EM_n^{(K)}$. As there is no code available online, up to our knowledge, we develop our own code considering the algorithm proposed in Chen and Li (2009). We did not recover the asymptotic distribution under H_0 claimed by the author; in particular, we noticed a higher estimated value for $P(EM_n^{(K)} \leq 0)$ and slightly different estimated type I errors. Using a similar setting used for M8, we noticed that good power is reached for obvious mixtures.

7.2 Conclusion

This paper brought a new insight towards the problem of testing the number of components of a mixture. We compare two approaches, one based on the Likelihood-Ratio Test, and the other on the EM-test; guidelines for practionners are summarized in the previous subsection, and an illustration of both apporaches using different models is proposed. Classical results obtained in the frame of Likelihood Ratio Test (Ghosh and Sen (1985), Garel (2001)) rely on the true value of the parameters under H_0. But this value is unknown and may be difficult to estimate in a general framework. Then, it would be tempting to consider that it is only a theoretical result, without possible application. Another issue could be considered. We need a procedure useful with real data coming from both hypotheses. Under H_0, the classical ML estimators could be used. Then, we would find a good α-level of the test; but, if we are under H_1, these estimators are dramatically inefficient. Therefore we have to decide what could be the H_0 component, being under H_1. A solution could be to consider the main component (the one with the largest weight) as the H_0 component and to use either an EM algorithm or a robust procedure to estimate this component. This could be investigated. Another surprising by-product is the inaccuracy of our simulation results with respect to the EM Test results in the case of (M9). We have no clear

explanation so that further investigations would be needed. Finally, we stress that this work allowed us to add a few contributions as codes implementing the LRT approach, that will be included in a futur update of the mixtools package (Benaglia et al., 2009) for the R statistical software (R Core Team, 2016).

References

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Bertillon (1874). *Démographie figurée de la France.* Masson, Paris.

Bertillon, L. (1876). *Moyenne. Dictionnaire encyclopédique des sciences médicales*, pages 296–324. MASSON, Paris.

Bhattacharya, C. (1967). A simple method for resolution of a distribution into its gaussian components. *Biometrics*, 23:115–135.

Bickel, P. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In et al., J. G., editor, *Statistics and Probability*, pages 83–96. Wiley Eastern Limited.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications.* Monographs on Statistics and Applied Probability 81. Chapman & Hall.

Charlier, C. (1906). Researches into the theory of probability. *Lunds Univ. Ars. N y foljd*, Afd. 2.1(5).

Charlier, C. and Wicksell, S. (1924). On the dissection of frequency functions. *Arkiv f. Mathematik, Astron. och Fysik*, Bd.18(6).

Chen, H. and Chen, J. (2001). Large sample distribution of the likelihood ratio test for normal mixtures. *Statistics and Probability Letters*, 52:125–133.

Chen, H., Chen, J., and Kalbfleisch, D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal Royal Statistical Society*, 63:19–29.

Chen, H., Chen, J., and Kalbfleisch, D. (2004). Testing for a finite mixture model with two components. *Journal Royal Statistical Society*, 66:95–115.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *The Annals of Statistics*, 37:2523–2542.

Chen, J. and Li, P. (2011). Tuning the EM-test for finite mixture models. *The Canadian Journal of Statistics*, 39(3):389–404.

Chen, J., Li, P., and Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107:1096–1115.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25:573–578.

Cohen, A. (1967). Estimation in mixtures of two normal distributions. *Technometrics*, 9:15–28.

Dacunha-Castelle, D. and Gassiat, E. (1997). Testing in locally conic models and application to mixture models. *Esaim Prob. Statistics*, 1:285–317.

Everitt, B. and Hand, D. (1981). *Finite mixture distributions*. Chapman and Hall, London.

Feng, Z. and McCulloch, C. (1996). Using bootstrap likelihood ratio in finite mixture models. *Journal of the Royal Statistical Society B*, pages 609–617.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer-Verlag, New-York.

Garel, B. (2001). Likelihood ratio test for univariate gaussian mixture. *Journal of Statistical Planning and Inference*, 96:325–350.

Garel, B. (2005). Asymptotic theory of the likelihood ratio test for the identification of a mixture. *Journal of Statistical Planning and Inference*, 131:272–296.

Garel, B. (2013). *Modèles de mélanges : le nombre de composants*, chapter 3, pages 57–84. Technip.

Garel, B. and Goussanou, F. (2002). Removing separation conditions in a 1 against 3-components gaussian mixture problem. In Jajuga, K., Sokolowski, A., and Bock, H.-H., editors, *Classification, Clustering and Data Analysis*, pages 61–73. Springer.

Ghosh, J. and Sen, P. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In Cam, L. L. and Olshen, R., editors, *Proc. Berkeley Conf. in honor of Jerzy Neyman and Jack Kiefer*, pages 789–806, Monterey, Wadsworth.

Gottschalk, V. (1948). Symmetric bimodal frequency curves. *Journal of the Franklin Institute*, 245:245–252.

Gridgeman, N. (1970). A comparison of two methods of analysis of mixtures of normal distributions. *Technometrics*, 12:823–833.

Hall, P. and Stewart, M. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 13:795–800.

Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures. In Cam, L. L. and Olshen, R., editors, *Proc. Berkeley Conf. in honor of Jerzy Neyman and Jack Kiefer*, pages 807–810, Monterey, Wadsworth.

Lemdani, M. and Pons, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli*, 5:705–719.

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105(491):1084–1092.

Li, P., Chen, J., and Marriot, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96:411–426.

Li, S., Chen, J., and Li., P. (2016). *MixtureInf: Inference for Finite Mixture Models*. R package version 1.1.

Lindsay, B. (1995). *Mixture models: theory, geometry and applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31:807–832.

Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal Statistical Planning and Inference*, 123:61–81.

Maciejowska, K. (2013). Assessing the number of components in a normal mixture: an alternative approach. Technical report, Munich Personal RePEc Archive, Wroclaw University of Technology, Poland, CERGE-EI, Prague, Czech Republic.

McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appli. Statist.*, 36:318–324.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Aplications to Clustering*. Marcel Dekker, New-York.

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley and Sons, New-York.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Newcomb, S. (1882). Discussion and results of observations on transits of mercury from 1677 to 1881. *Astr. Papers*, 1:363–487.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of mathematics*, 8:343–366.

Pearson, K. (1894). Testing homogeneity in a multivariate mixture model. *Philosophical Transactions of the Royal Society of London, A*, 185:71–110.

Quetelet, A. (1846). *Lettres à S.A.R. le Duc régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités appliquée aux sciences morales et politiques*. Hayez, Bruxelles.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, B*, 10:159–203.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for non identifiable distributions. *Ann. Statistics*, 9:225–228.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normal. *Journal of the American Statistical Association*, 89:1096–1102.

Schlattmann, P. (2009). *Medical applications of finite mixture models*. Statistics for Biology and Health. Springer-Verlag, Berlin, Heidelberg.

Thode, H., Finch, S., and Mendell, N. (1988). Simulated percentage points for the null distribution of the likelihood ratio for a mixture of two normals. *Biometrics*, 44:1195–1201.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester.

Wilks, S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62.

Wolfe, J. (1971). *A Monte-Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions*, volume 72-2 of *Technical Bulletin STB*. U.S. Nav. Pers. and Train. Res. Lab., San Diego.